Predicting Birth Trends in Colombia Post-COVID-19 Using Time Series Models

EPH31

Authors: Álamo, A.[1], Arciniegas, J.[1], Escobar, O.[1], La Rotta, J.[1], Reyes, J.M.[1]

[1]Pfizer, North Latin American Cluster.

OBJECTIVE

- Population projections estimated by official entities in Colombia did not consider the most recent trends and exogenous economics variables. This study aims to forecast the number of newborns in Colombia for the period of 2025-2029 using historical data and sociodemographic variables.

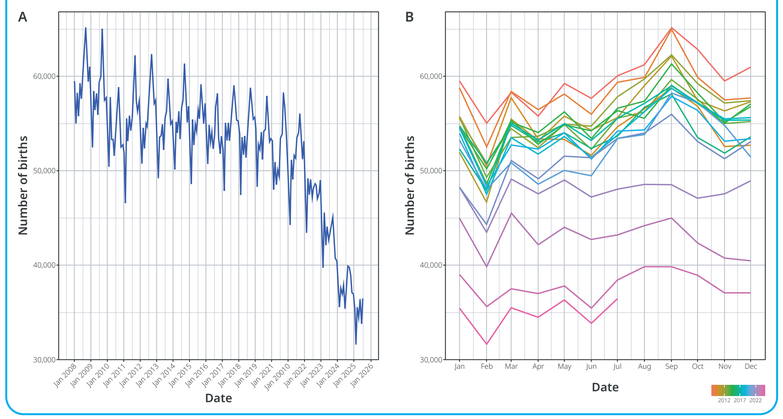
METHODS

- This analysis employed three statistical approaches: exponential smoothing (ETS and Theta models) and autoregressive integrated moving average models (ARIMA). The newborns monthly time series was extracted from the Colombian Statistics Bureau (DANE by its acronym in Spanish) available through the Integrated Social Protection Information System (SISPRO by its acronym in Spanish)[1].
- To test if unit root is present in a time series, the Augmented Dickey–Fuller (ADF), Phillips–Perron (PP), and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests were conducted.
- Considering the population dynamics framework and the social determinants of health, co-variables such as rate of women of childbearing age, unemployment rate, female net migration and covid period were included and were extracted from official entities[2–5].
- The series were segmented into a training dataset (June 2012 to July 2024) and a test dataset (August 2024 to July 2025). Shapiro-Wilk normality test was conducted to assess the residuals distribution and Ljung-Box test were used to detect serial correlation.
- Exploratory and descriptive analysis were made to characterize the series. To evaluate goodness-of-fit, metrics such as the RMSE, MAPE and the Winkler score were calculated. The computation of yearly forecasts and its prediction intervals (PI) were developed through bootstrap.

RESULTS

- The monthly birth time series points out the demographic shift currently underway in Colombia. Before 2010, the average monthly number of births was approximately 60,000; however, in recent years, this figure has decreased to around 40,000 births per month. Furthermore, the series demonstrates clear seasonality, characterized by an annual decrease in births during February and a peak in September. Notably, since 2022, these peaks have exhibited a tendency to become flattened (Figure 1).

Figure 1. Monthly births in Colombia, January 2008 – July 2025, A. time series and B. seasonal plot



- ADF, PP and KPSS tests indicated that the original series were non-stationary, while the differenced series satisfied the criteria for stationarity. Several transformations were analyzed (e.g. logarithmic transformation, differentiated and logarithmic differences), the differentiated series satisfied the stationarity tests (Table 1). Consequently, the modeling process was conducted using the differentiated series.

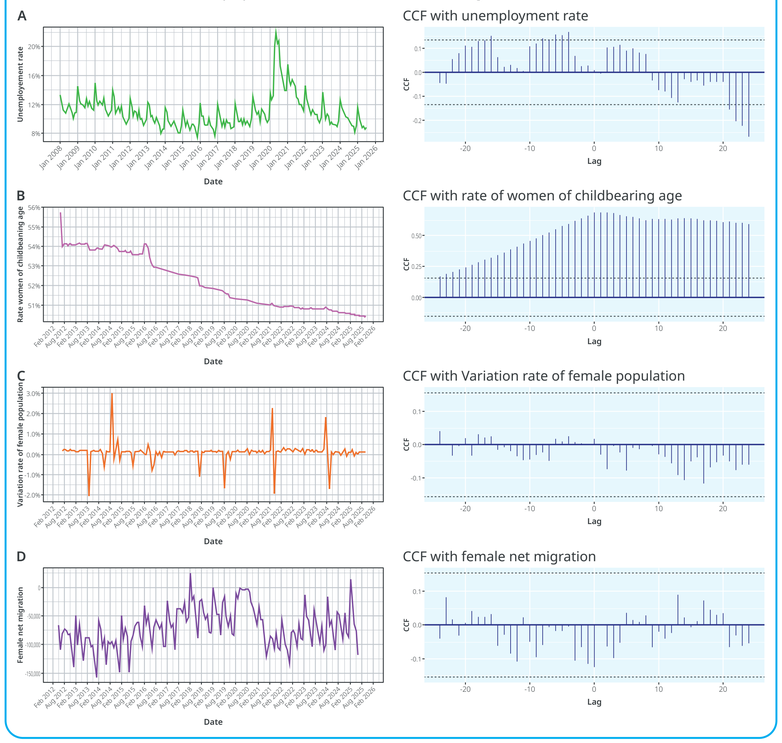
RESULTS (cont)

Table 1. Stationarity tests

Series evaluated	Test	Null hypothesis	Statistic	P-value
Births (original)	ADF	The series has a unit root (non-stationary)	-2.7642	>0.1
Births (original)	PP	The series has a unit root (non-stationary)	-47.536	<0.01
Births (original)	KPSS	The series is level stationary	2.8015	<0.01
Births (differentiated)	ADF	The series has a unit root (non-stationary)	-8.492	<0.01
Births (differentiated)	PP	The series has a unit root (non-stationary)	-273.36	<0.01
Births (differentiated)	KPSS	The series is level stationary	0.04039	>0.1

- The auxiliary variables presented different behaviors; the unemployment rate exhibited a seasonal trend, particularly pronounced in January each year; the proportion of women of childbearing age exhibited a consistent downward trajectory; the female net migration showed both an upward trend and seasonal variation; and the variation rate of female population showed remained stable, with fluctuations not exceeding 2% (Figure 2). The cross-correlation function (CCF) of these variables with births illustrates that only the unemployment rate and the rate of women of childbearing age show meaningful cross-correlations with the births series (Figure 2).

Figure 2. Auxiliary variables: A) Unemployment rate, B) Rate of women of childbearing age, C) Variation rate of female population and D) Female net migration



- A pseudo-dummy variable was created to account for the incidence that the COVID-19 pandemic might have introduced into the births time series behavior, taking values of 0 prior to the pandemic, 1 in January 2020 and a downward exponential trend since that period.
- By accuracy the best model was ETS with multiplicative errors, seasonality and damped additive trend (MAdM). RMSE/MAPE/MASE/Winkler scores were 518, 1.4, 0.2 and 4,401 (Table 2). Nevertheless, only ARIMA family models satisfied most of the residuals' assumptions, such as normality and the absence of serial correlation in seasonal lags as recommended by Mahdi[6].

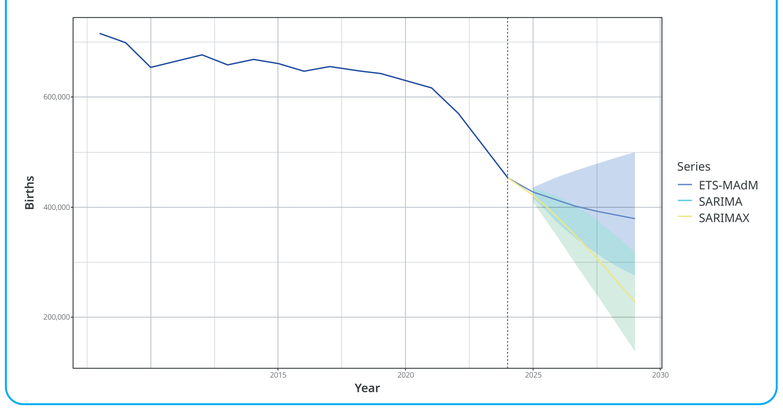
RESULTS (cont)

Table 2. Time Series Evaluation Metrics

Model	RMSE	MAPE	MASE	Winkler
ETS_MAdM	518.5	1.4	0.2	4,401.4
Theta_additive	1,052.7	2.9	0.5	6,187.9
Theta_multiplicative	1,209.8	3.4	0.5	5,893.6
ETS	1,371.1	3.8	0.6	6,762.6
SARIMA	2,105.4	5.9	0.9	7,404.0
SARIMAX	2,439.3	6.8	1.1	9,225.2

- The ETS_MAdM suggests an annual forecast for 2025 of 427,545 births (PI95% 418,179-436,397), with a decrease of 5.8% from the previous year. This trend continues through 2029, when 379,212 births were projected (PI95% 276463.3- 501294.1). A SARIMA model [p=8, d=1, q=1, D=1] forecasted 422,141 births for 2025 (PI95% 418,179- 436,397) and (PI95% 276,463-501,294) for 2029. Finally, the model that included the auxiliary variables, the SARIMAX model [p=8, d=1, q=1, D=1], forecasted similar numbers than the aforementioned model, with 422,120 births for 2025 (PI95% 409,050 – 435,105) and 227,180 births for 2029 (PI95% 137,976 – 315,707).

Figure 3. Annual births (2008-2024) and forecasts (2025 - 2029)



CONCLUSION

- Our model projected a decrease in the newborn population over the next 5 years but much sharper than official estimates. Exponential smoothing models presented the best fit and the most conservative values, while ARIMA models projected a steep drop in the expected number of births and the less accurate fit. Additional considerations may be warranted to the impact of social factors when making future population projections.

REFERENCE

1. DANE, MSPS-SISPRO. Cubo de estadísticas vitales: 2008-2025 preliminar n.d. https://www.sispro.gov.co/catalogos/Pages/catalogo-de-informacion.aspx (accessed October 10, 2025).
2. MSPS-SISPRO. Cubo de BDUA: Corte Noviembre 2025 2025. https://www.sispro.gov.co/catalogos/Pages/catalogo-de-informacion.aspx (accessed November 11, 2025).
3. DANE. Empleo y desempleo: Información a octubre 2025 2025. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo (accessed November 14, 2025).
4. Unidad Administrativa Especial Migración Colombia. Entradas de extranjeros a Colombia: corte a junio 2025. Datos Abiertos 2025. https://www.datos.gov.co/Estad-sticas-Nacionales/Entradas-de-extranjeros-a-Colombia/96sh-4v8d/about_data (accessed October 10, 2025).
5. Unidad Administrativa Especial Migración Colombia. Salidas de colombianos desde el territorio nacional: Corte a junio 2025. Datos Abiertos 2025. https://www.datos.gov.co/en/Estad-sticas-Nacionales/Salidas-de-colombianos-desde-el-territorio-naciona/efw5-jiej/about_data (accessed October 10, 2025).
6. Mahdi E. Portmanteau test statistics for seasonal serial correlation in time series models. SpringerPlus 2016;5:1485. https://doi.org/10.1186/s40064-016-3167-4.

Disclosures: This study was funded by Pfizer. Presented at ISPOR international 2026 in Philadelphia, PA, 17- 20 May.